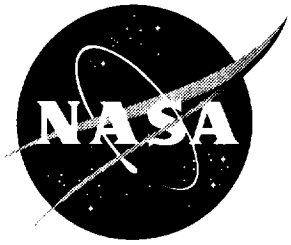
111 32
2112-1

NASA Technical Memorandum 112871



Lyceum: A Multi-Protocol Digital Library Gateway

Ming-Hokng Maa
Massachusetts Institute of Technology, Cambridge, Massachusetts

Sandra L. Esler
University of Alabama, Tuscaloosa, Alabama

Michael L. Nelson
Langley Research Center, Hampton, Virginia

July 1997

National Aeronautics and
Space Administration
Langley Research Center
Hampton, Virginia 23681-0001

Lyceum: A Multi-Protocol Digital Library Gateway

Ming-Hokng Maa
mmaa@mit.edu
Massachusetts Institute of Technology
Cambridge, MA 02139

Sandra L. Esler
sesler@eng.ua.edu
University of Alabama
Tuscaloosa, AL 354877

Michael L. Nelson
m.l.nelson@larc.nasa.gov
NASA Langley Research Center, MS 124
Hampton, VA 23681

Abstract

Lyceum is a prototype scalable query gateway that provides a logically central interface to multi-protocol and physically distributed, digital libraries of scientific and technical information. Lyceum processes queries to multiple syntactically distinct search engines used by various distributed information servers from a single logically central interface without modification of the remote search engines. A working prototype (http://www.larc.nasa.gov/lyceum/) *demonstrates the capabilities, potentials, and advantages of this type of meta-search engine by providing access to over 50 servers covering over 20 disciplines.*

Introduction

Internet document and information archival, indexing, and distribution is typically embodied in widely heterogeneous and distributed information servers, employing search engines with user interfaces and query syntax that vary significantly. The incompatibility and non-interoperability between search engines and the lack of a common and unified interface between users and distributed information servers present a significant challenge to the design of meta-search engines and the indexing and retrieval of comprehensive and apropos information on the Internet. Because of this growing problem we developed Lyceum (http://www.larc.nasa.gov/lyceum/), a working scalable query gateway meta-search engine that provides a common and unified interface to widely heterogeneous and distributed information servers.

Design

A number of digital libraries exist on the World Wide Web (WWW). However, many overlap with other information servers and are incomplete, both in terms of the content they provide and the subject areas they cover. This requires the user to have detailed knowledge of where the various digital libraries are and what resources can be found in them. In short, users must perform extensive integration of the information they receive from various sources (Figure 1). In addition, these information servers often utilize different search engines to index their information. Various search engines, e.g., Harvest [Bowman, et al., 1995], Wide Area Information Server (WAIS) [Kahle, et al., 1992], freeWAIS-sf [Pfeifer, et al., 1995]. often require different syntax for search functions such as Boolean searches, result limiting and ranking, case sensitivity, search method, and other miscellaneous return options. This variability has typically discouraged attempts to implement meta-search engines. We have previously examined various digital library architectures and have found the distributed architecture with the contributors being the authoring organization or individuals as the most desirable architecture (Esler & Nelson, 1997). There are two primary advantages to distributing information among multiple servers versus implementing one centralized information server:

1. Each information server is now responsible only for maintaining information local to an organization.
2. One canonical information server or database which covers the entire spectrum of scientific research is patently neither feasible nor desirable.

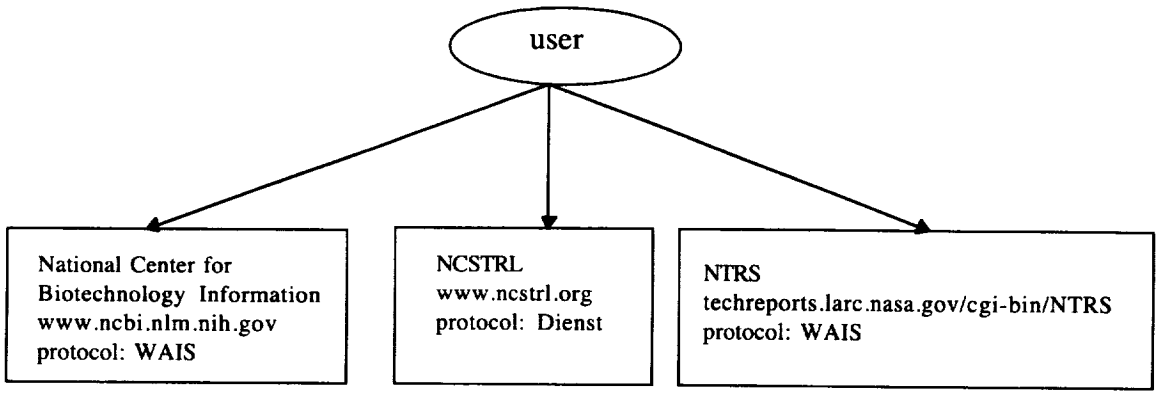


Figure 1: User Performs Multiple Searches

What is needed is a logically central interface to physically distributed and heterogeneous databases. As a gateway server, Lyceum allows users to query syntactically distinct search engines from a single interface by formatting a user's query submission to conform to the appropriate search engine query syntax and options (Figure 2).

Because of the rapid pace with which new information servers are established, Lyceum was designed for scalability. The ability to add both individual information servers and other meta-search engines similar to Lyceum enables Lyceum to access many already cataloged information servers by taking advantage of pre-existing services such as NCSTRL [Davis, et al., 1995], building on the work of others rather than recompiling databases of individual information servers. Within a distributed architecture, there are two methods to aggregating digital library resources:

1. Encourage the proliferation and wide spread adoption of a single digital library protocol
2. Provide a protocol conversion functionality to gateway between heterogeneous resources

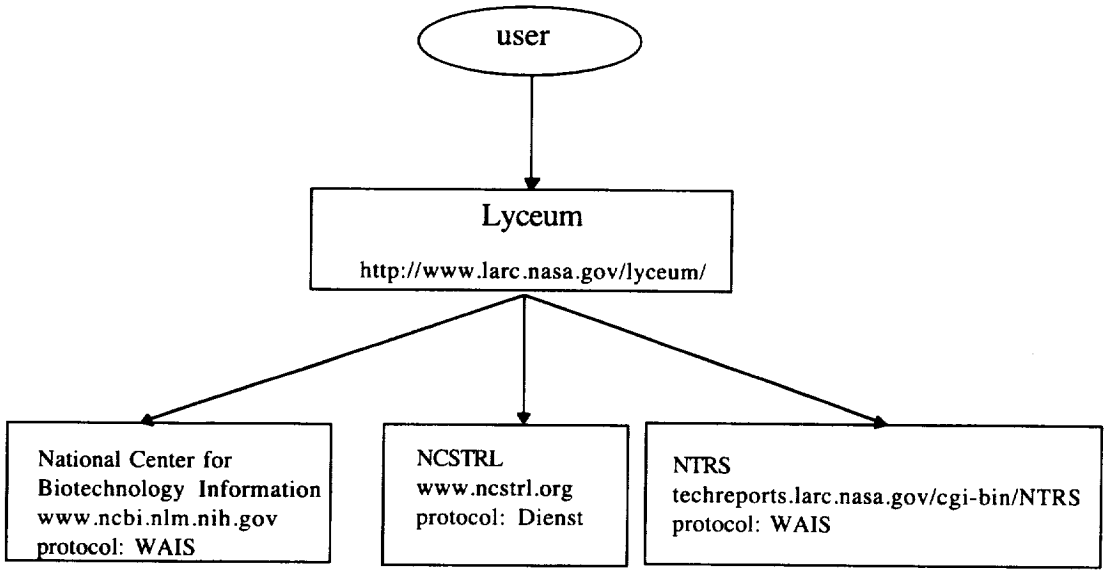


Figure 2: Lyceum Performs Multiple Searches for the User

The first method is what is employed by NCSTRL and NCSTRL+ [Nelson, et al., 1997]. The second method is that adopted by Lyceum. Finally, Lyceum was designed to require no special input or coordination from the remote information servers. Inclusion of and gatewaying to the remote information servers is performed without any action from the remote servers. Indeed, direct queries from users or from Lyceum are indistinguishable to the remote servers.

Implementation

Lyceum runs on a UNIX platform as a package of Perl Common Gateway Interface (CGI) scripts separated into client, server database, and administration scripts. Because Lyceum typically gateways to numerous information servers per query, Lyceum sends queries to information servers in parallel. This significantly reduces the user's idle time by shifting the time dependence of query results away from slow bottleneck servers and stalled connections towards a process where query results are returned as soon as they are available. In addition, this also prevents bad Internet connections from deadlocking the entire query. The Lyceum client, therefore, is a multi-forking client designed to gateway query requests in parallel. As users submit queries to Lyceum, the client forks and submits an individually tailored query (dependent on the requirements of each search engine) to each information server. When all requests have been forked, the client waits for the results of each query, post-formatting and merging the returned query results, and displaying the results to the user as they return. This implementation is similar to the parallel search algorithm implemented in the NASA Technical Report Server (NTRS) [Nelson & Maa, 1996].

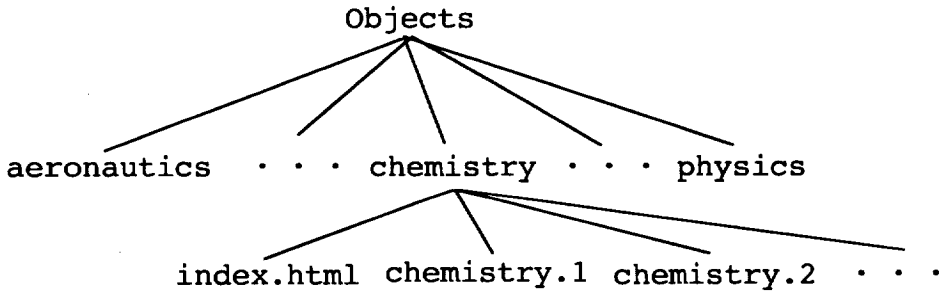


Figure 3: Objects Placed in the Subject Categories

Secondly, Lyceum was written with the goal of data type independence. Because Lyceum does not explicitly maintain records of the information type archived at various information servers, Lyceum may be used to index not only scientific information, but any arbitrary indexed information. To achieve data type independence, the server database, which maintains information on each of the information servers accessed by Lyceum, simply consists of a structured directory tree (Figure 3). Renaming the directory structure automatically changes the information type that users see. Each file in this directory structure is a record for an information server, containing relevant descriptors and most importantly, the unique uniform resource locator (URL) that is used to query the server's search engine. This form URL stores the syntax rules by which to format a user's query string for compatibility with a particular search engine. Figure 4 shows an example of the data files used to describe individual servers in Lyceum. Although new values can be added, the following values are currently used:

Type: WAIS, Glimpse, Split, Other
Category: aerodynamics, aeronautics, astronautics, biology, chemistry, computer, earth, economics, energy, engineering, environmental, geography, geology, hierarchy, materials, mathematics, medicine, meteorology, multi-categories, nonlinear, physical, physics, psychology, social, zoology
Sub-Category: Technical-Reports, Journals, Proceedings, Bibliographies, Newsletters, Other

```
Name:   Electronic Conference on Trends in Organic Chemisty (ECTOC)
Category:       chemistry
Sub-Category:   Proceedings
Description:
Home:   http://www.ch.ic.ac.uk/ectoc/
Date:   Fri Aug  2 10:11:40 EDT 1996
Form:   http://www.ch.ic.ac.uk/cgi-bin/pursuit-ectoc?$keywords
Type:   Other:+
Boolean:        YES
POC:
POCemail:
POCphone:
Comments:
```

Figure 4: File Format for an Object

Finally, Lyceum was designed to be as maintenance free as possible. All system administration takes place through a WWW interface, providing a flexible layer that shields the administrator from the programming code (Figure 5). All query interfaces are generated on the fly according to the contents of the server database, erasing the need for periodic updating of the server database. Because query interfaces are generated on the fly (Figures 6 & 7), delays slightly increase as the server database populates. Currently, the delay is minimal, but if Lyceum's population grew to the point where the dynamic construction of the interface became noticeable to the user, we could switch the interface construction to be static and periodically regenerated (e.g., every 12 hours). The search results interface is a concatenation of what the various servers would return if they had been searched individually. Typical examples are too lengthy to be placed in a typical figure, so the reader is encouraged to experiment.

Lyceum Administrative Interface

Add objects	*Add objects manually into the site database.*
Edit/Remove Objects	*Edits or removes individual objects in the database.*
Grant requests	*Grant previously submitted requests for object addition.*
Generate statistics	*Generate statistics about the site database and server workload.*
Set Search Defaults	*Set the default sites to be search*
Rotate Log	*Rotate old log file and create new log file.*
Change Password	*Change password for all system administration files.*
Flush Old Objects	*Flushes out the recover directory.*
Verify URL's	*Verify URL's in Database*

Figure 5: The Lyceum Administrative Interface and Functions

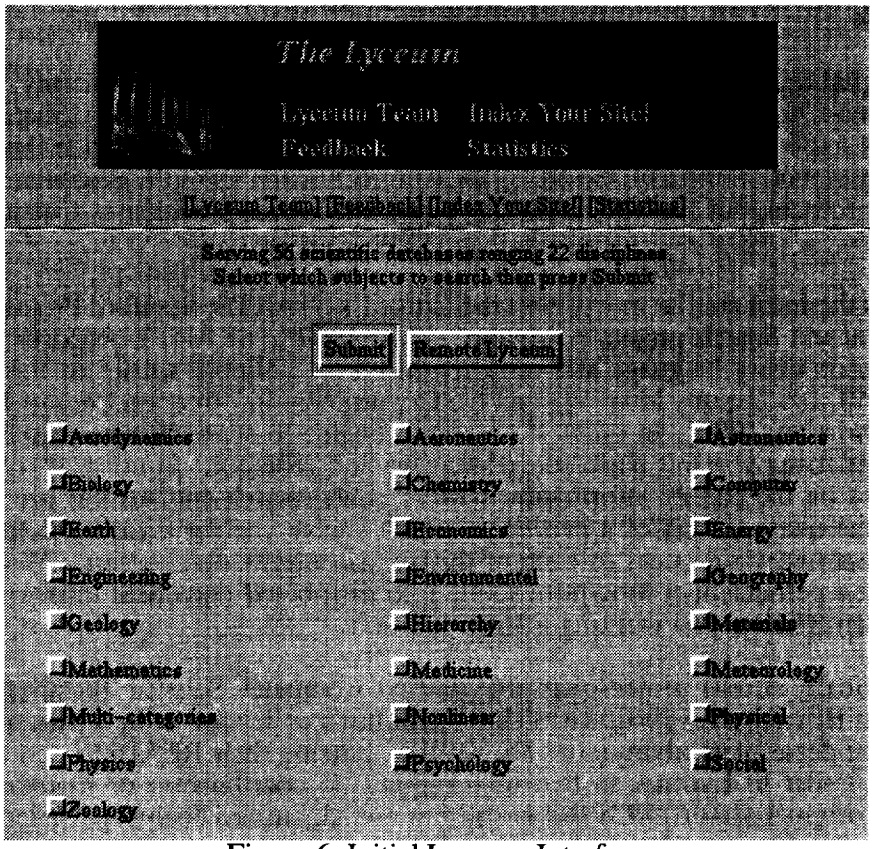


Figure 6: Initial Lyceum Interface

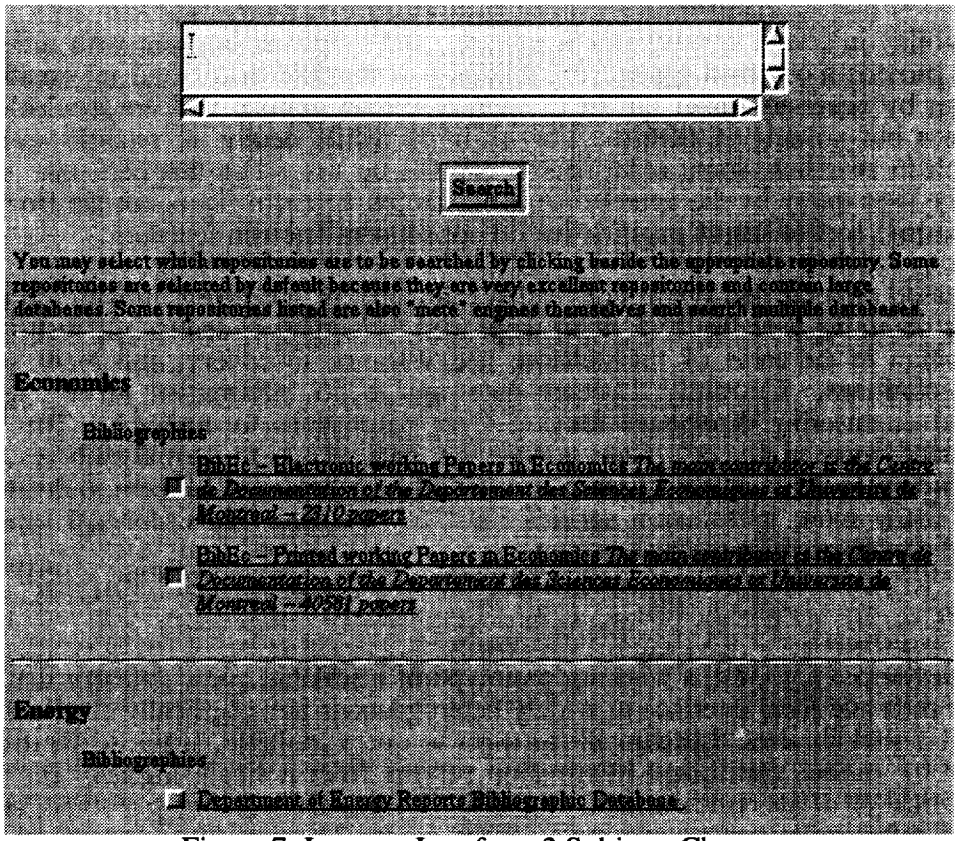


Figure 7: Lyceum Interface, 2 Subjects Chosen

Related Work

Few digital library projects attempt to build a multi-discipline digital library. The University of Illinois Urbana-Champaign's portion of the Digital Library Initiative (DLI) [Shatz, et al., 1996] does, but it has explicit agreements with various journal publishers to provide their titles using a homogeneous protocol. Similarly, the NCSTRL+ project builds its multi-discipline project using a homogeneous protocol, but has the authoring organizations as participants instead of traditional publishers.

Perhaps the most similar to Lyceum is Stanford's STARTS: Stanford Protocol Proposal for Internet Retrieval and Search project [Gravano, et al., 1997]. STARTS proposes to implement a minimally common query language and protocol based on a simple subset of the Z39.50-11995 type-101 [Z39.50, 1995] query language and the Government Information Locator Service (GILS) attribute set [Christian, 1996]. In order for a search engine to support meta-searching, it must be actively modified to support this basic query language and protocol. In contrast, Lyceum requires no modifications on the part of information servers or the search engines that are used. Rather, knowledge of the query syntax of a particular search engine enables Lyceum to format the query into an equivalent and valid query to each of the remote information servers. While this method may not provide a permanent or long-term solution to distributed document indexing and retrieval, it does provide an efficient and working interim solution.

A historically similar project was the Unified Computer Science Technical Report Index (UCSTRI) [Van Heyningen, 1994]. UCSTRI would use a collection of heustrics to index various computer science department anonymous ftp servers, parsing their README and abstracts files. UCSTRI was similar to Lyceum in that both require no coordination or modification with the provider of the original server. UCSTRI is still functional, but has been superseded by NCSTRL.

Discussion and Future Work

Variability in search syntax, search options, and output is inherent with different search engines. To provide a common gatewaying interface to multiple distributed information servers, Lyceum must by necessity either ignore or modify certain options that are available in certain search engines but missing in others. For example, while nearly all modern search engines provide Boolean searches, some older search engines do not. For this particular situation, in formatting the user query by the appropriate syntax rules, Lyceum filters out the Boolean syntax from query strings sent to search engines that do not support Boolean searches.

Increased protocol and syntax conversion is an area for improvement with Lyceum. Other areas include obtaining user feedback about improved formatting of search and results interfaces, more automation in the areas of maintenance and resource discovery, and more sophisticated cataloging techniques. Currently, Lyceum gateways to 56 information servers spanning 22 scientific and engineering disciplines ranging from aeronautics to zoology. The information servers included range in diversity from government data servers to journal archives. While this has been sufficient to demonstrate Lyceum conceptually, in order for Lyceum to demonstrate true application, many more information servers must be included. We encourage others to test, evaluate, and contribute resources to Lyceum.

Conclusion

We developed Lyceum, a working prototype of a scalable query gateway that provides a logically common and local interface to widely heterogeneous and physically distributed scientific and technical digital libraries. Lyceum allows users to query multiple, syntactically distinct search engines used by various distributed information servers from a single logically central interface without modification of the remote search engines. The current working prototype, incorporating more than 56 information servers and meta-servers across 22 scientific and technical disciplines, demonstrates the capabilities, potentials, and advantages of this type of meta-search engine. Suggestions and contributions to Lyceum are welcome. Contact the authors for more information.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	July 1997	Technical Memorandum

4. TITLE AND SUBTITLE
Lyceum: A Multi-Protocol Digital Library Gateway

5. FUNDING NUMBERS

6. AUTHOR(S)
Ming-Hokng Maa, Sandra L. Esler, Michael L. Nelson

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
NASA Langley Research Center
Hampton, VA 23681-2199

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
National Aeronautics and Space Administration
Washington, DC 20546-0001

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
NASA TM-112871

11. SUPPLEMENTARY NOTES
Ming-Hokng Maa, Massachusetts Institute of Technology, Cambridge MA; Sandra L. Esler, University of Alabama, Tuscaloosa, AL; Michael L. Nelson, NASA Langley Research Center, Hampton, VA

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Unclassified–Unlimited
Subject Category 82 Distribution: Nonstandard
Availability: NASA CASI (301) 621-0390

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)
Lyceum is a prototype scalable query gateway that provides a logically central interface to multi-protocol and physically distributed, digital libraries of scientific and technical information. Lyceum processes queries to multiple syntactically distinct search engines used by various distributed information servers from a single logically central interface without modification of the remote search engines. A working prototype (http://www.larc.nasa.gov/lyceum/) demonstrates the capabilities, potentials, and advantages of this type of meta-search engine by providing access to over 50 servers covering over 20 disciplines.

14. SUBJECT TERMS	15. NUMBER OF PAGES
WWW, Digital Libraries, STI, Distributed Information Retrieval	9
	16. PRICE CODE: A02

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	

References

Bowman, C. Mic; Danzig, Peter B.; Hardy, Darren R.; Manber, Udi ; Schwartz, Michael F. & Wessels, Duane P. (1995). Harvest: A Scalable, Customizable Discovery and Access System. University of Colorado Computer Science Technical Report CU-CS-732-94.

Christian, Elliot (1996). GILS: What is it? Where is it going?, D-lib Magazine, December 1996, *http://www.dlib.org/dlib/december96/12christian.html*

Davis, James R.; Krafft, Dean B.; & Lagoze, Carl (1995). Diesnt: Building a Production Technical Report Server. *Advances in Digital Libraries*, Springer-Verlag, pp. 211-222.

Esler, Sandra L. & Nelson, Michael L. (1997). The Evolution of Scientific and Technical Information Distribution, *to appear in the Journal of the American Society for Information Science*.

Gravano, L.; Chang, C.-C. K.; Garcia-Molina, H.; Paepcke, A. (1997). STARTS: Stanford Proposal for Internet Meta-Searching, *Proceedings of the International Conference on Management of Data (SIGMOD)*, May 12-15, Tuscon, AZ.

Kahle, Brewster; Morris, Harry; Davis, Franklin; Tine, Kevin; Hart, Clare & Palmer, Robin (1992). Wide Area Information Servers: An Executive Information System for Unstructured Files. *Electronic Networking: Research, Applications, and Policy*, 2(1), pp. 59-68.

Nelson, Michael L. & Maa, Ming-Hokng (1996). Optimizing the NASA Technical Report Server, *Internet Research: Electronic Networking Applications and Policy*, 6(1), pp. 64-70.

Nelson, Michael L.; Maly, Kurt; Shen, Stewart N. T. (1997). Buckets, Clusters, and Dienst. Old Dominion University Computer Science Technical Report 97-30 and NASA TM-112877.

Pfeifer, Ulrich; Fuhr, Norbert; & Huynh, Tung (1995). Searching Structured Documents with the Enhanced Retrieval Functionality of freeWAIS-sf and Sfgate. *Computer Networks and ISDN Systems*, 27, pp. 1027-1036.

Schatz, B.; Mischo, W. H.; Cole, T. W.; Hardin, J. B.; Bishop, A. P. & Chen, H. (1996). Federating Diverse Collections of Scientific Literature. *IEEE Computer*, 29(5), pp. 28-36.

Van Heyningen, Marc (1994). The Unified Computer Science Technical Report Index: Lessons in indexing diverse resources. *Proceedings of the Second International World Wide Web Conference*, Chicago, IL, October 19-21, 1994, pp. 535-543.

Z39.50-1995 Maintenance Agency (1995). *http://lcweb.loc.gov/z3950/agency/1995doce.html*